Synopsis

Proteins are complex three dimensional molecules with a unique structure, which is responsible for its stability and function It is often stated that the folding of proteins to its unique native structure is the second genetic code However, the factors responsible for the formation and stabilization of these native protein structures are still not completely understood It is well known that analysis of stable folded three dimensional structures of proteins can provide insights into their folding, stability and function With the initiative of structural genomics, novel ways of analyzing the protein three dimensional structures are required to handle large-scale structure analysis We explore a few such methods based on principles of network and graph theory in this thesis so as to understand protein folding and function We already know that the non-covalent interactions like the hydrogen bonds, salt-bridges, hydrophobic interactions and van der Waals interactions play a significant role in stabilizing the protein tertiary structure Hence, we feel that representing the protein structure as a network of non-covalent interactions amongst the amino acid residues and analyzing the overall properties of such networks can provide insights into folding and stability of protein structures This thesis is essentially an attempt in this direction

A network or a graph is defined as a set of nodes connected by a set of edges Recently, novel methods of analyzing large-scale graphs have been elucidated on networks like the world-wide-web, internet, power grid, social and ecological networks, communication networks, etc (Barabasi, 2002) These studies have further been extended onto biological networks like the metabolic networks, transcription regulatory networks and protein interaction networks leading to a new and exciting area of research in biology called as network biology (Barabasi, 2002) These investigations have provided some interesting results on the properties of these self-organized real-world networks They have shown that many of these networks are non-random in nature and show a power-law distribution of the degree of the nodes Such networks are called scale-free networks They are characterized by the presence of a small number of highly connected nodes called hubs which impart stability and

robustness to these networks from random attacks (Barabasi, 2002) The presence of such hubs also reduces the distance between any two nodes in these networks due to which they are also referred to as small-world networks

An interesting application of the network concepts is in the understanding of protein structures The representation of protein structures as graphs and the analysis of their graph properties have already been carried out by various groups to compare protein structures, understand protein stability, folding and function and even in comparative modeling (Vishveshwara *et al*, 2002) The large-scale network parameters and the small-world network properties of the protein structure networks have also been investigated In all these studies, the definition of nodes and edges in the protein structure networks and the method of analysis vary according the aim of the study In this thesis, we have used different definitions of nodes and edges to obtain protein structure networks and have analyzed them from a very novel perspective We have also used an already established graph spectral cluster determining algorithm (Kannan & Vishveshwara, 1999) to obtain clusters of amino acid residues in protein structures and protein-protein interfaces We have addressed several relevant problems in structural biology using this network approach and have obtained insights into the factors affecting protein stability and protein-protein interactions The main aim of this thesis is to apply the simple concepts from graph theory onto large number of protein structures so as to shed light onto problems related to protein stability and folding Further, this thesis presents a simple and elegant method to analyze large-scale protein structures to obtain biological insights We also highlight the role of non-covalent interactions in protein stability, folding and interactions

This thesis has eight chapters Chapter one is the introduction chapter, which provides a general introduction and the relevant literature in the field of structural and computational biology It also provides an overview of the concepts of networks and graph theory and their applications in biology The present status of the applications of the network concept to protein structure analysis is also provided Chapter two explains in detail the methodology used in the analysis of protein structures carried out in this thesis. A brief survey of the concepts of graph theory and graph spectral theory is provided. The methods of constructing protein structure graphs and parameters used in analyzing them are explained in detail in this chapter

Protein structures are represented as a network of non-covalent interactions between the amino acid residues, with the strength of the non-covalent interaction being an important parameter in the graph construction Thus, we get a weighted representation of the protein structure graph Methods of identifying amino acid clusters and hubs in protein structures are also clearly explained

Chapter three applies the network concepts onto 232 globular monomeric proteins The protein structure graphs of these proteins generated as given in the methods chapter and the overall network properties are analyzed in detail The topological behavior and the clusters and hubs in these protein structure graphs are also analyzed Interestingly, we find that the protein structure graphs, unlike many other self-organized networks, show a complex topology, with the network organization depending on the non-covalent interaction strengths used in the network construction Moreover, a transition is observed at a critical strength of interaction in all the proteins, as monitored by the size of the largest cluster (giant component) in the graph Amazingly, this transition occurs within a narrow range of interaction strength for all the proteins, irrespective of the size or the fold topology. Further, the preferences of the amino acids to form hubs in the protein structures also vary according to the interaction strengths used The hubs are also found to play a role in integrating different secondary structures in the tertiary structure of the protein These protein structure network concepts are then applied onto the structure graphs of a set of ten thermophilic proteins and their mesophilic counterparts so as to understand the factors imparting additional stability to the thermophilic proteins This study clearly shows that the network representation of protein structures and network parameters used in their analysis are able to account for the additional stability of the thermophilic proteins and aid in identifying residues that can alter the thermal stability of proteins (Brinda & Vishveshwara, 2005a)

Chapter four is an extension of the analysis carried out on monomeric proteins in chapter three onto a large set of oligomeric proteins The oligomeric structure graphs of these proteins are constructed and analyzed in detail Specifically, the strong interface clusters and hubs detected in these oligomers are identified to be important for the stability of the interface The network parameters of the oligomeric graphs are compared with those of the monomeric protein structure graphs Topologically, the oligomeric protein structure networks are very similar to the

monomeric networks However, the interfaces of the oligomers show differences in the amino acid hub preferences Many protein interfaces are found to be stronger than their monomeric protein cores based on the non-covalent interaction strengths and sizes of the interface clusters We find that the interface hubs identified in fifteen protein complexes chosen from the Alanine scanning energetics database (ASEDB), correlate very well with experimentally determined binding energy hot spots (Brinda & Vishveshwara, 2005b)

Chapters three and four have dealt with the overall network properties of monomeric and oligomeric protein structure graphs, with special emphasis to the hub concept In chapters five, six and seven, we deviate from the general network aspects discussed in chapters three and four, giving more emphasis to the spectral parameters of protein structure graphs and the clustering algorithm based on the spectra of protein graphs To understand the principles of protein-protein association, we require an in-depth analysis of amino acid networks at protein interfaces Hence, in chapter five, we provide a rigorous analysis of the interface clusters obtained in a set of homodimeric proteins The interface clusters in these proteins are identified using a graph spectral method and the residues important for maintaining the integrity of the interfaces are identified using a combination of graph spectral parameters, accessible surface area calculations and conservation across proteins from different species This analysis identifies residues in these proteins, which correlate well with experimentally determined hot spots at their interfaces Thus, we have two robust methods of identifying hot spots on protein interfaces, as explained in chapters four and five We further obtain a method for predicting interacting surfaces on monomers based on the results from this analysis (Brinda *et al* , 2002).

Chapters six and seven handle specific problems in protein structural biology from a graph spectral perspective In chapter six, we analyze a set of proteins called legume lectins to understand the factors affecting their quaternary association These proteins are known to have very similar sequences and tertiary structures However, they have very different modes of quaternary associations. We have used the graph-spectral cluster determining algorithm along with traditional multiple sequence alignment methods to obtain sequentially and structurally conserved residues in the interfaces of these proteins These residues are found to impart specificities to the modes of quaternary associations in the proteins Such an analysis also has a high

predictive value since it is able to predict the mode of association in legume lectins with unknown structures (Brinda *et al*, 2004) This study is extended onto other lectins sharing the jelly-roll fold like galectins and pentraxins to provide a comprehensive analysis on the factors influencing oligomerization in these lectins (Brinda *et al*, 2005)

In chapter seven, we provide a novel and elegant method to identify structural domains in proteins using concepts from graph spectral theory It considers both covalent and non-covalent interactions of amino acid residues in the graph representation The sub-clusters identified in the connected protein graph using a graph-spectral method are partitioned into structural domains Two different representations using protein backbone and side-chain interactions are used for the identification of domains. Both methods give robust results and correlate well with already existing methods of domain assignment in proteins The advantage of this method is that it involves a single numeric computation and is completely automated This is of significance since such objective automated methods of domain assignment are required to handle the large scale output from structural genomics Moreover, this method also provides information on the residues forming the domain-domain interface, which are otherwise very difficult to obtain (Sistla *et al*, 2005)

Chapter eight provides the summary and conclusions of the studies presented in this thesis In general, the thesis has explored ways of analyzing protein structures by representing them as networks and applying the concepts from network biology and graph theory These have been done efficiently to answer several important questions related to protein structure, stability and interactions Many of the network concepts and analysis protocols presented in this thesis have very high predictive value and hence can be a very useful tool for protein structure analysis The studies carried out in this thesis and their results can motivate further theoretical and experimental studies, a brief account of which is also provided in chapter nine

There are two appendices in the thesis, which elaborate two small pieces of work unrelated to the main focus of the thesis In appendix 1, we come up with some generic geometrical parameters that take into account the constraints offered by the backbone conformation of non-covalently interacting residues in protein structures, so as to understand the factors leading to the folding and stabilization of protein structures In appendix 2, we try to address the problem of obtaining a better scoring

function for the sequence alignments of remotely related protein structures using a perceptron-based learning algorithm Both these problems deviate significantly from the main aspects of this thesis and hence have been presented in the Appendix section

References

1 Barabasi A L 2002 Linked The new science of networks Persues Publishing, Cambridge, Massachusetts

2 Vishveshwara S , Brinda K V. and Kannan N 2002 Protein structure Insights from graph theory *J Th Comp Chem* , **1(1)** 187-211

3 Kannan N and Vishveshwara S 1999 Identification of side-chain clusters in protein structures by a graph spectral method *J Mol Biol.*, **292(2)** 441-64

4 Brinda K V and Vishveshwara S 2005a A network representation of protein structures Implications to protein stability Accepted in *Biophysical Journal*

5 Brinda K V and Vishveshwara S 2005b Oligomeric protein structure networks insights into protein-protein interactions Submitted to *BMC Bioinformatics*

6 Brinda K V , Kannan N , Vishveshwara S 2002 Analysis of homodimeric protein interfaces by graph-spectral methods *Protein Engineering*, **4**, 265-77

7 K V Brinda, Nivedita Mitra, Avadhesha Surolia and Saraswathi Vishveshwara 2004 Determinants of quaternary association in legume lectins *Protein Science*, **13**, 1735-1749

8 Brinda K V , Surolia A & Vishveshwara S 2005. Insights into the quaternary association of proteins through structure graphs A case study of lectins *Biochemical Journal*, In press.

9 Ramesh K Sistla, Brinda K V and Saraswathi Vishveshwara 2005 Identification of Domains and Domain Interface Residues in Multidomain Proteins from Graph Spectral Method, *Proteins Struct Funct Bioinfo* **59(3),** 616-626